Detection and Classification of Breast Cancer Based-On Terahertz Imaging Technique Using Artificial Neural Network & K-Nearest Neighbor Algorithm

Hassan Jassim Motlak
Department of Electrical Engineering,
University of Babylon, College of Engineering, Babylon, Iraq.

Orcid: 0000-0001-7798-0893

Safa Isam Hakeem
Department of Electrical Engineering,
University of Babylon, College of Engineering, Babylon, Iraq.

Abstract

Terahertz imaging system has emerged as a powerful platform for a wide range of applications, especially in the security and medical systems. Because of its nonionizing nature. Terahertz imaging system is a save and non-invasive medical imaging method for detecting cancer and provides an internal contrast between the natural and the cancerous tissue because it is sensitive to the water content in the tissue. The proposed system includes the detection and classification of the abnormality for breast human tissues using terahertz imaging system. This system based-on an Artificial Neural Network (ANN) & K-Nearest Neighbor (KNN) algorithm that classifies breast cancer as either Benign or Malignant through pattern recognition. Terahertz images were preprocessed at different stages including (resize image, median filter and segmentation). The second step is the important information was extracted from these images in vector format by using texture features (GLCM) and invariant moments algorithm. Finally classification of the normality and abnormality of breast tissues are done using ANN and KNN algorithm. Simulation results show that the proposed system is efficient and accurate for detection and classification of Breast Cancer. Where the value of accuracy obtained by ANN is equal to 98.2% and 96.4% obtained by KNN algorithm. Moreover, the processing time of classifier is less than one second.

Keyword: Terahertz Imaging technique, Texture features (GLCM), Invariant moment , Classifier algorithms ANN and KNN.

INTRODUCTION

Cancer is one of the main causes of death worldwide, surpassing only heart disease at present, the accounts of Cancer refer to nearly one of every four deaths. For some types of cancers, the early detection can decrease the mortality cases. The common techniques of cancer diagnostic in the imaging base technologies are biopsy, MRI and X-rays and T-ray. The process of breast cancer screening has done through several literatures, [16] L. Álvarez Menéndeza is tested the employment of the Support Vector Machines (SVM) of the polynomial kernel of in the breast cancer diagnosis. He investigated the performance of these models to classify breast cancer. [17] M. Noorazlan,M. Mohd and Mohd Muzafar [2011] image processing was used to detect breast cancer in the early stages. X-ray technology was used. These images have been processed using the image row These processes include optimizing and filtering images and determine texture feature extraction using Gabor Wavelet algorithm. Anthony j. Fitzgerald (2012) [17] used data reduction techniques to help disaggregate the data of the terahertz pulse obtained from the tumor and normal breast tissue. The data reduction can be classified into three methods which used the principal components of the pulses, and ten inference parameters. The vector support machine combines with radial basis function technique was used to perform the classification based on the parameter ten components, this method give 92% accuracy. It was show that the results under sum classification and controlled condition give better accuracy for cancer classification and normal tissue of breast. Mohamed Azhardeen.S, Usha.S [2014] [18] used to detect breast cancer by using Super Vector Machine (SVM) classifier , the detection of the cancer follows , preprocessing , feature extraction using symlet wavelet and classification.

In this paper Artificial Neural Network (ANN) & K- Nearest Neighbor (KNN) algorithm were used to detect and classify breast cancer images obtained from terahertz imaging technique. Comparison between the proposed other literatures are obtained. Simulation results showed efficient results with less consumption time of the breast cancer detection using

ANN algorithm. The simulation was carried out using Intel ® core ™ i5-2520M CPU @ 2.50 GHz PC with 4GB RAM, windows 7, 64-bit OS, and MATLAB 2014a software.

TERAHERTZ IMAGING TECHNIQUE

Over recent years, the technologies of terahertz (THZ) system which represents optically-driven technologies are dramatically expanding and developing. The moneymaking applications of these technologies are now beginning to become known. THz radiation is non-ionizing and have least effects on the human body , It has very large absorption due to water and Metals highly reflect terahertz radiation, combining the terahertz and the pattern recognition which related to multivariate statistical tools, lead to potentially provide a rapid and non-invasive method to diagnose and detect diseases [1], the radiation of terahertz in medical system are more accurate in diagnosis THZ technology allows high resolution surface of tissue and provide good contrast between several of (TPI) human soft tissue, the difference between cancer and normal tissues in terahertz absorbance using in the pathologic diagnosis [2], Getting images and spectral data is not main difficulties in the terahertz technologies which used as a medical diagnostic tool, but getting helpful information from these data is the main difficulty, Studies that are used MRI and PET show that tumors increase the water content. Because of the water nature which has clear resonance in the terahertz band, the imaging in this region is highly sensitive to the concentration of water content, this lead to Increase image contrast in the terahertz, There is a need to identify multivariate statistical tools of the pattern recognition that are used in analysis and classification of this data. The pattern classification is the technique which assigns raw data to one of several categories, typically raw data is preprocessed before being provided to a classifier.

ARCHITECTURE OF THE PROPOSED SYSTEM

The proposed system includes various methods of the THZ image analysis compared to the previous researches, THZ Images pass through important stages of processing, It does not take long, as well as processing speed, due to the development of image processing.

During the imaging process, not all images give a direct notion to the doctor, for this reason image processing techniques are important in this case, to clarify the patient's condition as well as the amount of abnormality , type of disease, …etc.

Figure (1) consists of stages including reading T-ray image and then processed these Images into several stages including resize image ,filter, k-mean segmentation and extract meaningful information from image called the process of extraction features in two ways such as gray level co-occurrence matrix and invariant moments, These features are considered as inputs to the Neural network and k-nearest neighbor classifiers in order to classify breast cancer into normal and abnormal samples

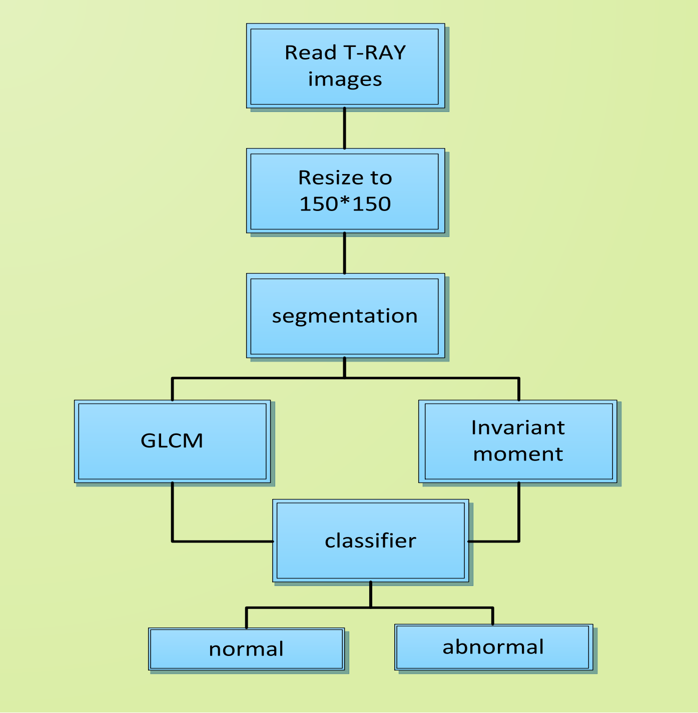

Figure 1: Block diagram of the proposed system for breast cancer detection and classification

DATA DESCRIPTION

(Normal and abnormal images of breast tissue), Images are taken from twenty female with (rang equal to 36-72 years), figure (2) shows normal and abnormal images of the breast tissue [3].

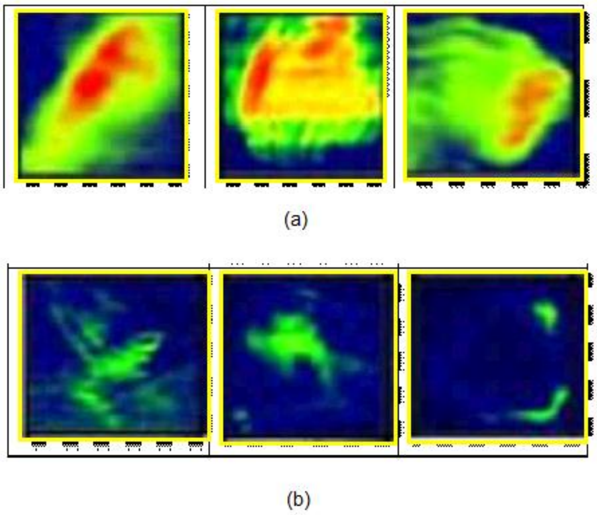

Figure 2 : Terahertz Images of the breast cancer tissues (a) abnormal images and (b) normal images .

PRE-PROCESSING

Due to the increased need for advanced image processing, it requires processing of images in a primary way called pre-processing, These processes include eliminating noise from images by using filters, re-sizing images with suitable dimensions, Convert images from colored to gray as well as how to preserve the edges of the image, there are many ways to improve images depending on the type of imaging, [4] and these processes are :

- Read image: image is read by using this command
- B = imread('breast1.jpg');
- Resize image: %resize images to 150*150*3 unit 8 color image
- J = imresize(B, [150 150]);
- Filtering : use median filter to smooth and remove noise from each channel
- filter each channel separately
- Segmentation used k-mean algorithm:

Stages of k-means include Guess the number of cluster values denoted by k, choose the centers of clusters randomly, find the cluster center or mean , find the distance between each cluster center and each pixel , where if the distance is close or near to the center then move theses pixels to that cluster otherwise moved to the other cluster.

FEATURE VECTOR EXTRACTION ALGORITHMS

Means computing a descriptor from the pixels around each interest point. The simplest descriptor is just the raw pixel values in a small patch around the interest point. More sophisticated descriptors include SURF, HOG, FREAK, GLCM and invariant moments.

a. Texture features (GLCM)

Defined as the visual property or feature that indicate to the properties of the inherent surface of an pattern or object and its rapport with the besetment environment. Co-occurrence matrix has proven to be a common for extracting the texture feature in a gray scale image to be very effective. RGB image is converted to the gray scale image. a second-order method for creating texture features is provided by using co-occurrence probabilities [5].

co-occurrence probabilities represent the conditional joint probabilities of the all conjugal structures of the spatial window that have gray levels of interest given two parameters: : inter pixel space (δ) , orientation (θ). the probability is defined as:

$$p_r(x) = \{c_{xy} | (\delta, \theta)\} \tag{1}$$

where *Cxy* (probability of co-occurrence matrix between gray levels *x* and *y*) is defined as:

$$c_{xy} = {}^{p_{xy}}\big/_{\sum_{x,y}^{k} p_{xy}} \tag{2}$$

Where k is defined the number of gray levels , P_{xy} : the number of occurrences of the grey levels in the presented window, given the certain (*δ, θ*) couple, and The total in the denominator thus illustrate the total No of the grey level pairs (*x, y*) inside the window. Texture features are generated by applied statistics to the co-occurrence probabilities as shown in the following equations:

$$\text{Energy} = \sum_{x=1}^{k} \sum_{y=1}^{k} P_{xy}^2 \tag{3}$$

$$\text{Contrast}= \sum_{x=1}^{k} \sum_{y=1}^{k} (x-y)^2 \ . P_{xy} \tag{4}$$

$$\text{Entropy}= -\sum_{x=1}^{k} \sum_{y=1}^{k} Pxy \ \log_2 P_{xy} \tag{5}$$

$$\text{Correlation} = \sum_{x=1}^{k} \sum_{y=1}^{k} xy \ P_{xy} - \mu x . \mu y / \sigma x \ \sigma y \tag{6}$$

the conformable occurrence matrix was get from a size of 256 x 256. and grey scale quantification should be make ,the statistical properties such (contrast, energy, entropy and correlation) are determined in order to find the picture content. As shown in the following steps:

1. RGB image is transferred to the gray scale and then the picture co-occurrence matrix is found by using equations 1 and 2.
2. Contrast, energy, entropy and correlation are determined by using equation (3-6) in the four orientations or directions such as $o^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$, totally 16 texture features are found.
3. final texture features represented by mean and variance of the above four parameters and represented by the following vector:

T=μenergy,μContrast,μEntropy, μCorr,σEnergy,σContrast,σEntrop,σCorr.

Where

Entropy use to measure the randomness of a gray-level distribution, Energy use to measure the number of repeated pairs, Contrast use to measure the local contrast of an image., Homogeneity use to measure the local homogeneity of a pixel pair., Correlation provide a correlation between the two pixels in the pixel pair [5].

b. Invariant Moment Feature Extraction

Invariant moments are defined by Hu in 1962 , A given moment is a certain weighted average (moment) of the pixel density of the image, or the function of those moments, and usually chooses to have some attractive or interpreted property. It is widely used in the field of image processing and computer vision for the purpose of the shape recognition.

It is very useful to describe objects after segmentation, It is also used to extract the simple characteristics of the image, which include total intensity, centroid, and information about the orientation of the image

The first order moment will give you the center of mass, where the mass of a pixel is its intensity, second order moment will tell you how this mass varies around the center of mass, etc. In the same way as you obtain a frame of inertia for a real world object, you can obtain one from the image moments. That will give you the principal axes of the shape you want to describe. Shortly, the spatial moments give information about the object in the image [30], i.e. related (dependent) on the object position, these invariant moments are :

i. **Geometric Moments**

ii. Central Moments

iii. **Scale Invariant Moments**

iv. Rotation Invariant Moments

ARTIFICIAL NEURAL NETWORK ARCHITECTURE

Pattern recognition classifier is a network of two feed forward layers, with the sigmoid hidden neurons in the hidden layer and softmax output neurons in the output layer(pattern net) shown in figure (3) that can classify vectors arbitrary well, given sufficient neurons in its hidden stratum, the network will be train by using scale conjugate gradient back propagation (trainscg),the performance is evaluated using cross-entropy and confusion matrix, the input data set (feature vector) are divided into 70% training set and 30% testing set where(training sets are presented to the network during training ,and the network is adjusted according to its error) and testing sets is measure the network performance during and after training.

1. Q input vectors and Q target vectors are arranged as columns in a matrix form.
2. The input vectors and target vectors are divided into 70% training data set and 30% testing data set .
3. Set 10 neurons in the hidden layer, 2 neurons in the output layer, which is equal to the number of numbers in the target vector.
4. Train the network using (trainscg) command.
5. Measuring accuracy by using confusion matrix.
6. Save net

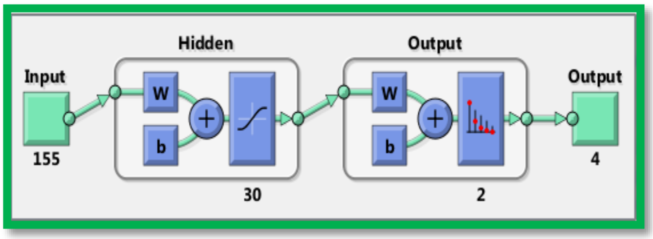

Figure 3 : Two layer feed forward network(pattern net)

K-NEAREST NEIGHBOR (KNN) ARCHITECTURE ALGORITHM

Is a very simple algorithm that stores all possible class cases, and works to classify new cases based on the functions of the distance (similarity measure) , KNN in the field of statistical estimation and pattern recognition.

The class is classified by the most votes of its neighbors, with the case being assigned to the most popular class among its K nearest neighbors measured by the function of the distance, If the value of k is one, the class is simply assigned to the nearest neighbor class, and classification is easy.

K-nearest neighbor algorithm steps:

1. Find the number of nearest neighbors which is known as k.
2. the distance between all the training samples and the query instance must be calculated.
3. The distance must be stored and determine closest or nearest neighbors based on the minimum distance.
4. Gather the Y class from the nearest neighbors.
5. Use the easy majority of the nearest neighbor class as the prediction No. for the query instance.[14]

SIMULATION RESULTS

FFNN classifier code that used to classify breast cancer data, Class 1 represents "+ive" and class 0 represents "-ive" for breast disease, after preprocessing we got 56 instances ready to use for training and testing the neural network. Training dataset contains 45 instances i.e (80% for trianing) and testing dataset contains 11 instances (20% for testing). The results of breast cancer detection are shown in the table (1).

Table 1: The classification results for breast cancer tissue using ANN

No	Feature extraction	TP	TN	FP	FN	%accuracy Over all accuracy	Performance ANN
1	4Glcm + seg.	41	13	2	0	96.4	0.3342
2	4Gclm + resize	41	14	1	0	98.2	1.9941e-04
3	Invariant+ seg	40	13	2	1	94.6	5.9105e-08

The above table shows the highest sensitivity and specificity were obtained when using texture features (GLCM) on the resized image and hu's invariant moment on the segmented image**.**

The performance results for breast cancer tissue using pattern net classifier are shown in table (2)

Table 2: Performance results for breast cancer tissue using ANN

No	Feature extraction	Gradient	Epoch	Best performance
1	4Glcm + seg.	8.08e-07	72	1.21e-07
2	4Gclm + resize	6.81e-07	56	1.25e-07
3	Invariant+ seg	8.51e-07	49	6.85e-08

Table 3: Test data classification using confusion matrix

No	Feature extraction	TP	TN	FP	FN	%accuracy
1	4Glcm + seg.	8	2	1	0	90.9%
2	4Gclm + resize	6	5	0	0	100%
4	Invariant+ seg	8	3	0	0	100%

Table (2) shows the results of performance during the process of classification with a different epoch in each case of features.

From the performance plot, the best performance is equal to **1.25e-07** when using texture feature extraction and it is equal to **6.85e-08** when using invariant moment feature extraction.

In breast cancer images there are few training samples, so it will be just one epoch is enough, but if the training package is huge, the training package should be divided into batches and for more than one epoch.

Table (3) shows test data classification for different feature vector with 30 neurons in the hidden layer, The accuracy of the classification was very good, Where the test samples are equal to 11 (8 abnormal samples & 3 normal samples of breast tissue) this indicates the accuracy of those features and it will facilitate any other classifier to easily distinguish abnormal breast tissue from normal breast tissue. Performance is plotted during the simulation it is showed that we obtain good performance at small number of iteration. Figure 4 shows the performance plot for (a) texture features (b) invariant moment features.

the mean square error dynamics for all data set in logarithmic scale error always decreasing , the performance plot shows perfect training at epoch 56.

Samples were also classified into normal and abnormal cases using k-nearest neighbor classifier as shown in table below :

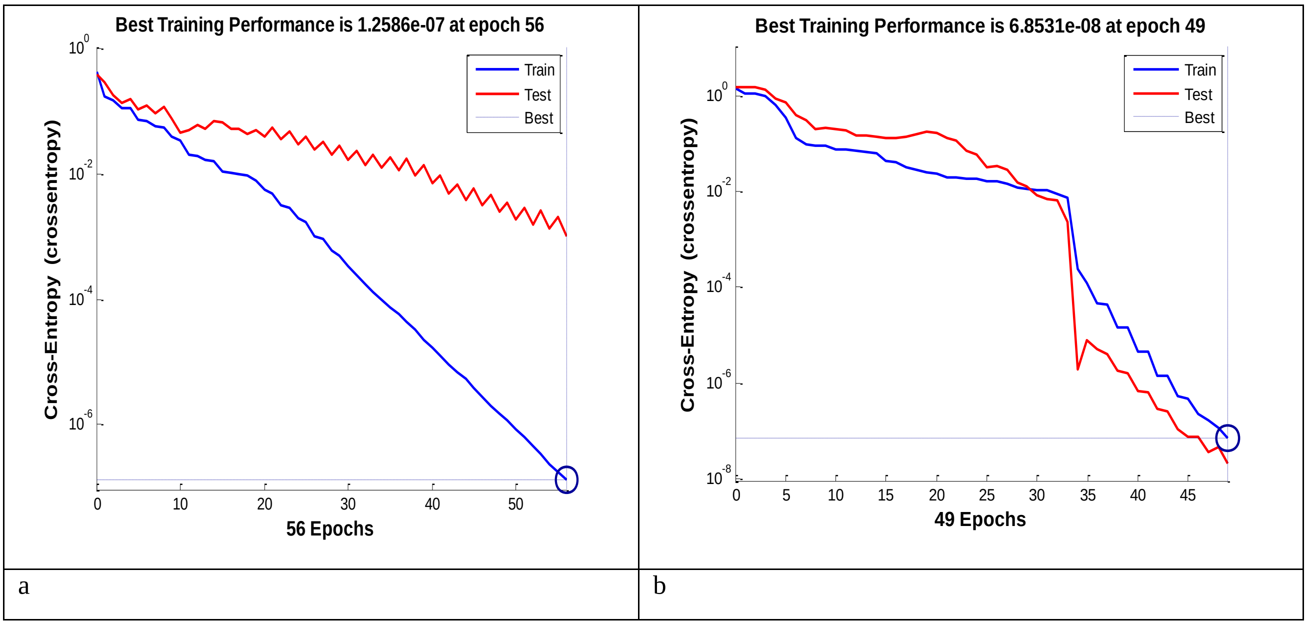

Figure 4: Performance plot of (a) texture features and (b) invariant moment faetures

Table 4: Classification results for breast tissue samples using KNN classifier

No	Feature extraction	TP	TN	FP	FN	% Over all accuracy	Error rate KNN
1	4Glcm + seg.	39	10	5	2	87.5	0.125
2	4Gclm + resize	39	15	0	2	96.4	0.035
4	Invariant+ seg	39	14	1	2	94.6	0.053

Table (4) shows classification results for breast tissue samples using KNN classifier , The best results were similar to the results of FFNN classifier when using the same features (i.e texture and invariant moment).where the overall accuracy is equal to 96.4 & 94.6 respectively. and the values of sensitivity and specificity are equal to (95.12 & 100) respectively**.**

Figure 5. Shows simple comparison between ANN and KNN classifier at 3 feature extraction methods (1) represent texture features at resized image (2) represent texture features at segmented image, (3): represents invariant moments, where texture features achieved the best accuracy at ANN&KNN classifier at method 2 with accuracies equal to 98.2 &96.4.

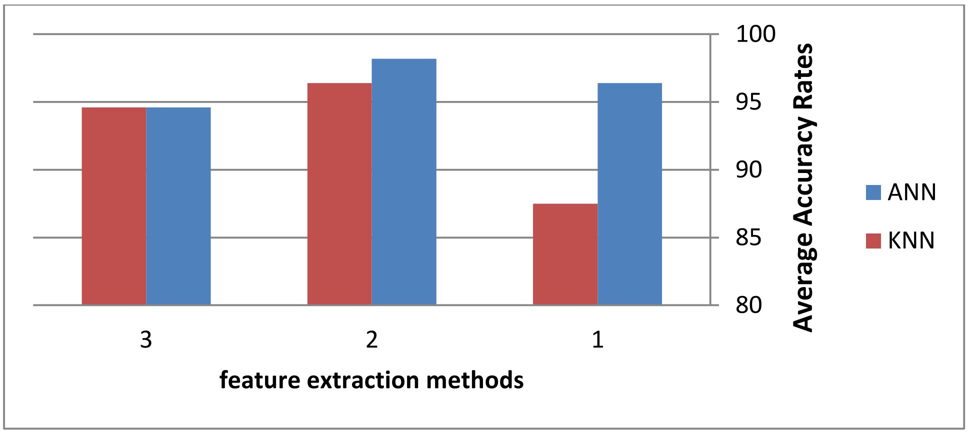

Figure 5: Comparison charts between ANN & KNN classifier of breast cancer tissues

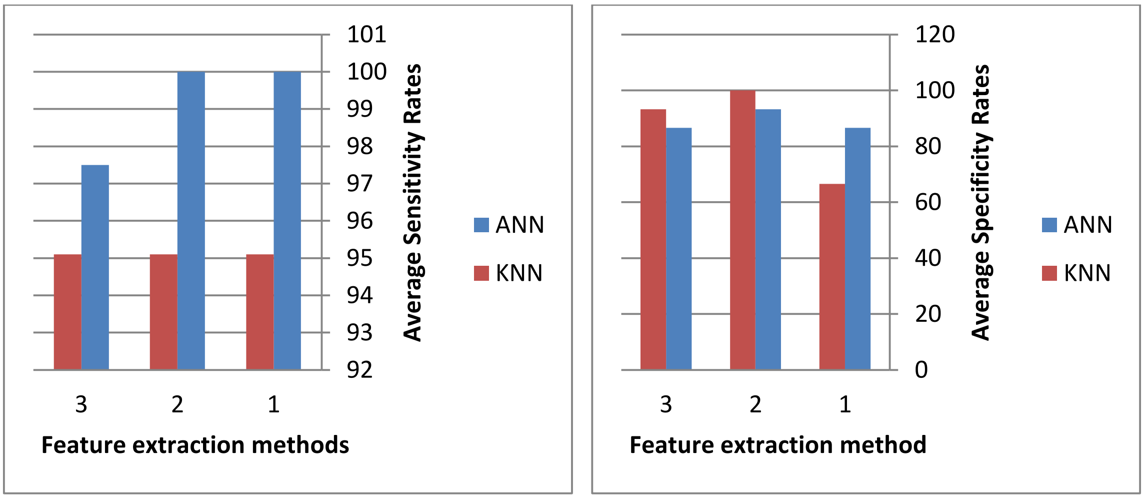

Figure 6: Average rate of Sensitivity and Specificity chart

Figure 6 above represents the best performance of the system in terms of specificity and sensitivity at texture features.

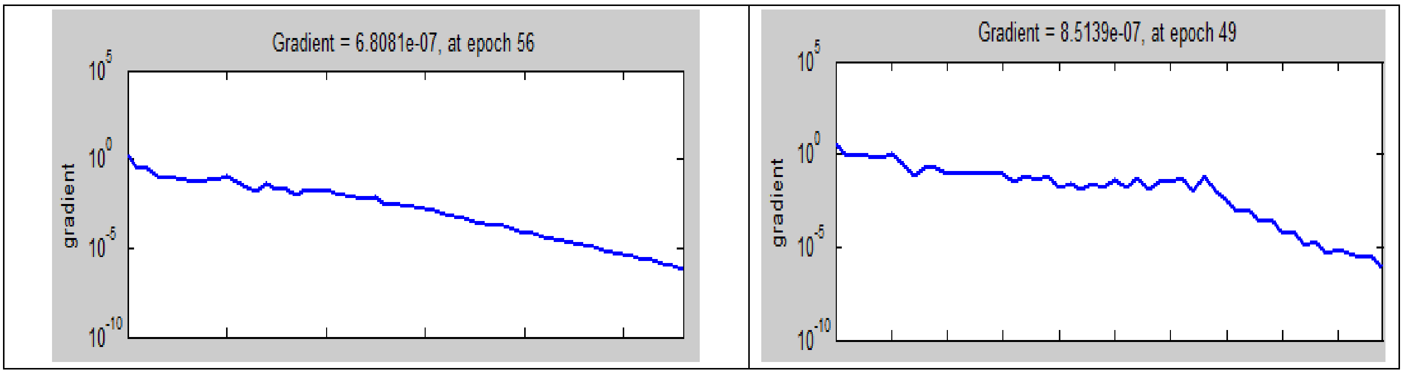

Figure 7: Pattern net parameter plots of GCLM with ANN

Figure 7 shows the best The performance measures in terms of accuracy. Texture and invariant moments in conjunction with ANN is used for the classification of breast tumor images, The best result in the training of the network is achieved after 49 iterations, as mean square error (performance) is -1.7e-07.

COMPARISON OF PRIOR AND CURRENT RESULTS

Breast cancer detection for many various parameters (texture , invariant moments, ANN & KNN) get the performance for each case and compare current results with other researchers such as ref.[8].

It has been shown from the comparison that the use of ANN and ANN algorithms is more efficient and very accurate at texture and invariant features and better than the super vector machine by 8 %.

CONCLUSION

The process of detecting breast cancer is a very difficult and time-consuming process. For helping pathologists by providing second opinions and reducing workload Terahertz technology has been used as a suitable tool for cancer detection but the disadvantage of applying this technique is consuming considerable time in manual diagnosis. The type of image by the radiologist is therefore used in the automatic detection and classification Artificial Neural Network and the nearest neighbor algorithm. The typology can greatly improve the diagnostic process in terms of accuracy and time requirements by automatically distinguishing between benign and malignant patterns. In this paper texture feature and invariant moment algorithms are used to detect breast cancer with two classifiers are ANN and KNN. The results shows excellent results of detection and classification in case invariant moment with ANN, where the accuracy is 98.2%.

Table 5: Breast cancer detection comparison results

parameters	Simulation results				Ref [8]		
No of samples	56	56	56	56	51	51	51
Feature extraction algorithm	GLCM	Invariant moments	GLCM	Invariant moments	ten heuristic parameters	PCA of the pulses	PCA of ten pram
Imaging technique	TPI	TPI	TPI	TPI	TPI	TPI	TPI
Classifier algorithm	ANN	ANN	KNN	KNN	SVM	SVM	SVM
sensitivity	100%	100%	95.1%	95.1%	80.3%	91.9%	90.3%
specificity	100%	100%	100%	93.3%	56.3%	91.8%	92.1%
accuracy	100%	100%	96.4%	94.6%	70.2%	92%	92%

REFERENCES

[1] T. Rainsford, S.P. Mickan and D. Abbott " T-ray sensing applications: review of Explosives and Weapons," in Proceedings of SPIE vol.5649, Bellingham, WA, , 2005, pp. 826-838.

[2] Wallace VP, Taday PF, Fitzgerald AJ, Woodward RM, Cluff J, et al "Terahertz pulsed imaging and spectroscopy for biomedical and pharmaceutical applications" world Journal of Radiology, vol.3. no. 3, March 2011, p.p.55-65.

[3] Hua Chen, Wen-Jeng Lee, and others, “Performance of THz fiber-scanning near-field microscopy to diagnose breast tumors” , Optics Express-, vol. 19, no.20, 2011, p.p. 1-10.

[4] Rafael C. Gonzalez and Richard E. Woods, "digital image processing using mat lab", Gates mark Publishing Library of Congress Control Number: 2009902793 , second edition, 2009.

[5] D. Chandrakala and S. Sumathi " Image Classification based on Color and Texture features using FRBFN network with Artificial Bee Colony Optimization Algorithm", International Journal of Computer Applications, vol.98, no.14 , July 2014 , p.p. 19-29.

[6] Paul L. Rosin, "Shape description by bending invariant moments", Computer Analysis of Images and Pattern journal, 2011 , p.p. 253–260.

[7] Philip C. Ashworth, Emma Pickwell-MacPherson, Elena Provenzano, Sarah E. Pinder and Anand D. Purushotham, " Terahertz pulsed spectroscopy of freshly excised human breast cancer", Optics Express,, vol.17 , no.15, 2013, p.p. 12444-12454.

[8] Fitzgerald AJ, Pinder S, Purushotham AD, O’Kelly P, Ashworth PC, "Classification of terahertz-pulsed imaging data from excised breast tissue", Journal of Biomedical Optics, vol. 17, no.1, January 2012 , pp.1-10.

[9] A. Mahmoud and U. Binti Obaidellah ," Artificial Intelligence Techniques for Cancer Detection and

Classification", European Scientific journal , vol.13, no.3, January 2017, p.p .342-370.

[10] Carlos Gershenson, "Artificial Neural Networks for Beginners", available from website> http://arxiv.org/ftp/cs/papers/0308/0308031.pdf, 6/7/2015,5:39 > Accessed 2003.

[11] Samuel H. Huang and Hong-Chao Zhang, " Artificial Neural Networks in Manufacturing: Concepts, Applications, and perspectives", IEEE Transactions on Components Packaging and Manufacturing Technology, vol.17. no.2 Part A · July 1994, p.p.212-228.

[12] Ajith Abrahem," Handbook of Measuring System Design", Third Edition, Stillwater, Ok, USA ,2005.

[13] Xue-Bin LI and Xiao-Ling YU, "Influence of Sample size on prediction of Animal Phenotype Value Using Back-Propagation Artificial Neural Network with Variable Hidden Neurons", The International Conference on Computational Intelligence and Software Engineering. Danvers: IEEE Computer Society, 2009: 257–260.

[14] Kardi Teknomo , "KNN numerical example hand computation", evoledu, available from website > http://people.revoledu.com/kardi/tutorial/KNN/KNN_Numerical-example.html> Accessed September 2017.

[15] Chandra Prasetyo Utomo, Aan Kardiana, Rika Yuliwulandari, " Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques ", International Journal of Advanced Research in Artificial Intelligence, vol. 3, no. 7, 2014, p.p. 10-14.

[16] Xiao-Xia Yina,∗, Sillas Hadjiloucasb, Yanchun Zhanga, Min-Ying Suc, Yuan Miaod, Derek Abbotte, " Pattern identification of biomedical images with time series: Contrasting THz pulse imaging with DCE-MRIs" , Artificial Intelligence in Medicine ,vol.67, 2016, pp. 1–23.